31 | 规划部署数据中心要考虑哪些重要因素?

2020-02-05 庄振运

性能工程高手课 进入课程 >

讲述：庄振运
时长 16:13　大小 13.00M

你好，我是庄振运。

上一讲我们讲了服务器的设计和部署，今天我们就来聊聊一个轻松的话题，一起来看看数据中心的秘密。

你们公司肯定有很多服务器。根据公司和服务的规模，可能有几十台甚至几万台服务器，大些的公司甚至会达到几百万台服务器。那么这么多的服务器都放在哪里呢？它们的家，就是数据中心。

你平时可能和服务器打交道比较多，离“数据中心”就比较遥远。但是我可以肯定地说，数据中心的知识，与我们每个 IT 从业人员（尤其是对于运维和性能工程的人员）是非常相关

的。

数据中心长什么样?

要讲数据中心，你可能会问数据中心到底长得什么样子?

其实许多科幻片里都会有类似数据中心的场景出现。一个数据中心往往有好几个大楼，每个大楼建筑物内部，都有着很好的空调和通风设施。大楼内一般分成几个数据大厅（Data Hall），每个大厅都有一排排的机架，中间留出足够的通道，方便数据中心的技术人员进行维护。

这是机架的背面。你可以看到，机架的布线必须非常整齐，太乱的话，日常中是很难进行维护的。

数据中心的规划和部署

对一个有全球用户的大互联网公司而言，它的容量通常也需要部署在全球范围内。运行公司各种服务的服务器，就放置在全球的数据中心里面。数据中心的建造成本很高，周期也较

长，也有足够的复杂度，所以中小公司，往往会租赁别人的数据中心的空间，来放置自己的服务器，甚至是直接租赁服务器。

但对于大公司，比如 Google、Facebook、Amazon、Microsoft、阿里巴巴、腾讯等，它们规模很大，大到不能靠租赁来运营。而且因为规模大，自己建造数据中心，从经济上看更加划算。比如亚马逊，就自己建造了很多数据中心。这些数据中心分布在全球各地，如下图所示。

之所以要在全球范围内建造数据中心，主要是为了性能，而不是为了节省成本。因为一个服务如果有全球的用户，这些用户就都需要和公司提供的服务快速交互，比如上传照片，播放视频等等。对于在全球范围内运行服务的公司而言，“数据中心离用户距离近”就是唯一的选择。

那么公司需要建多少数据中心呢？这个数量问题是容易解决的，就是**根据公司的规模和实际的需求，并且适当的做一些预测和远景规划**，而这就是我们下一讲会专门讨论的容量规划和预测。

数据中心要建在哪里呢？表面上看，这个问题也容易回答。前面说过，全球建造数据中心的初衷，就是让它们靠近客户，那当然是根据客户的地点来建造数据中心了，哪里有客户就把数据中心建在哪里。

这样的回答，道理上没错，只是考虑得不够全面。数据中心的选址还需要考虑很多因素，比如电力供应的稳定性、自然灾害发生情况、社会稳定性、所在国法律、人力资源、容量供应、建造成本等等。

这些因素都很容易理解，不过有意思的是，“所在国法律”是其中非常重要的一个因素。

一个公司总会存储各种用户数据，而公司是需要保护用户隐私的。但是很多国家的法律要求，建造在本国的数据中心，必须允许本国政府访问这些用户数据。这就与公司应尽的职责构成了冲突。

我们有时候开玩笑说，放眼全球，还真找不到几个国家，能够在该国不用担心警察会突然破门而入，用枪指着头，强迫数据中心员工交出客户的数据。如果考虑诸多这些因素，地球虽

大，却也难找到合适的地方建造数据中心。

数据中心建造还有一个特点，就是建造周期很长，从选址、规划，一直到建造完成，最少也需要好几年。所以，为了让一个数据中心能够长时间有效使用，建造一般也会刻意地分期完成。比如假设一个数据中心，最终会建造 6 个大楼，公司通常会分成 3 个阶段，一次建造两个大楼。

数据中心内服务器的生命周期

我们上一讲讨论了服务器的设计。一种服务器设计完成后，公司就可以部署了，你也需要对这个部署过程有个了解。对于一台服务器，它的生命周期经过 4 个阶段，包括购买和运送、按服务分配、运行管理、最终退休。

购买和运送

公司给了预算并且确定数据中心有了放置的空间和计划，就可以订购服务器了。服务器一般都是按照机架的单位批量购买。之所以要提前做好购买计划，是因为从购买到运送，一般需要几个月的时间。

服务分配

服务器放置在预定义的机架位置后，需要给它们通电。通电后，机架会自动在资产跟踪系统中注册。然后预配操作将安装操作系统以及许多其他软件。新服务器安装完成后，一般是放到备用池等待分配。服务所有者会提出申请服务器的要求，然后负责分配的团队，按照要求将容量分配给他们。

运行管理

在服务器的整个生命周期中，服务器和机架可能都需要进行维护。维护可能需要置换有故障的磁盘、更换坏的主板、重新启动服务器、重新安装 / 升级 / 修补操作系统、运行修复软件、诊断软件等。

我们通常的服务器设计，实际上已经考虑到了很多可能的维护工作。例如，既然换出磁盘是常见的修复任务，那么服务器设计可以让更换磁盘非常容易，无需任何工具比如螺丝刀等。

光荣退休

机架和服务器在数据中心的使用寿命是多久呢？通常约为 3-4 年。之后，我们需要让它光荣退休，将其从数据中心中移除，擦除所有数据，切碎磁盘，并遵循所有硬件的报废流程。

为什么需要让服务器及时退休？因为当超过使用寿命时，它们的组件就容易发生故障。在处理维修单和更换这些组件方面，会给数据中心的技术人员带来负担。同样，各个组件的保修期也可能快过期，甚至可能因为技术的发展，置换部件已经很难买到了。更重要的是，新的服务器替换旧的，也可以提高性能和效率，比如每一代的 CPU 都会比上一代更加强大。

Facebook 数据中心的网络部署

讲完了数据中心的架构和服务器的生命周期，我们再看看数据中心的网络部署。

数据中心内部的服务器之间，以及用户和服务器之间有大量的数据交换，这些对数据中心的网络部署也提出了很高的要求。为了能够更好地扩展，现代数据中心的网络设置也在不断地演化。下面我就用 Facebook 来举例，看看现代数据中心的情况。

Facebook 的生产网络本身，就是一个大型分布式系统，包括边缘网络、骨干网和数据中心内部网络。Facebook 的网络基础架构也在不断扩展。从 Facebook 到 Internet 的流量，我们称其为“**机器到用户**”的流量，非常庞大，并且还在不断增加。但是，这种流量，相对于数据中心内部发生的“**机器到机器**”流量，就只是冰山一角了，后者是前者的百倍以上。而且这种流量的增长速度，几乎每年都增长一倍。由此，我们也可以看出数据中心内部网络的重要性。

我们公司以前的数据中心网络，是使用集群 Cluster 构建的。集群是一个大型部署单元，涉及数百个服务器机柜，这些机柜的顶部（TOR）交换机，聚集在一组大型的交换机上。但是这种以集群为中心的体系结构有很大的局限性。

所以我们的新一代数据中心网络设计，就不是基于集群的，也就是说，不是按层次分配的集群系统。我们将网络分解为多个小的相同单元，也就是服务器 Pod，而不是大型集群，并在数据中心的所有 Pod 之间，创建了统一的高性能网络连接。

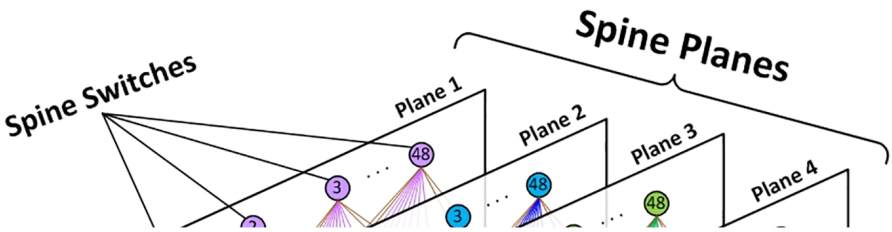
这里的 Pod，只是我们新架构中的标准“**网络单元**”。每个 Pod，由一组称为设备交换器的四个设备提供服务，从而可以根据需要进行扩展。当前的多数机架顶部交换机具有 4 个

40G 上行链路，为被连接的服务器提供 160G 的总带宽容量。下图就展示了一个 Pod 和 48 个机架的网络连接。

每个 Pod 的大小都一样，都是 48 个机柜，所以只需要基本的中型交换机就可以支持。对于机柜交换机的每个下行链路端口，我们在 Pod 的交换矩阵交换机上，保留相同数量的上行链路容量，这使我们能够**将网络性能扩展到在统计上无阻塞的水平**。

为了实现大楼范围的连通性，我们数据中心内部，创建了四个独立的骨干交换机“平面”，每个平面，可以最多扩展 48 个独立设备。

每个 Pod 的各个交换矩阵，都连接到其本地平面内的主干交换机，共同构成一个模块化的网络拓扑，能够容纳成千上万个连接 10G 的服务器。如下图所示：



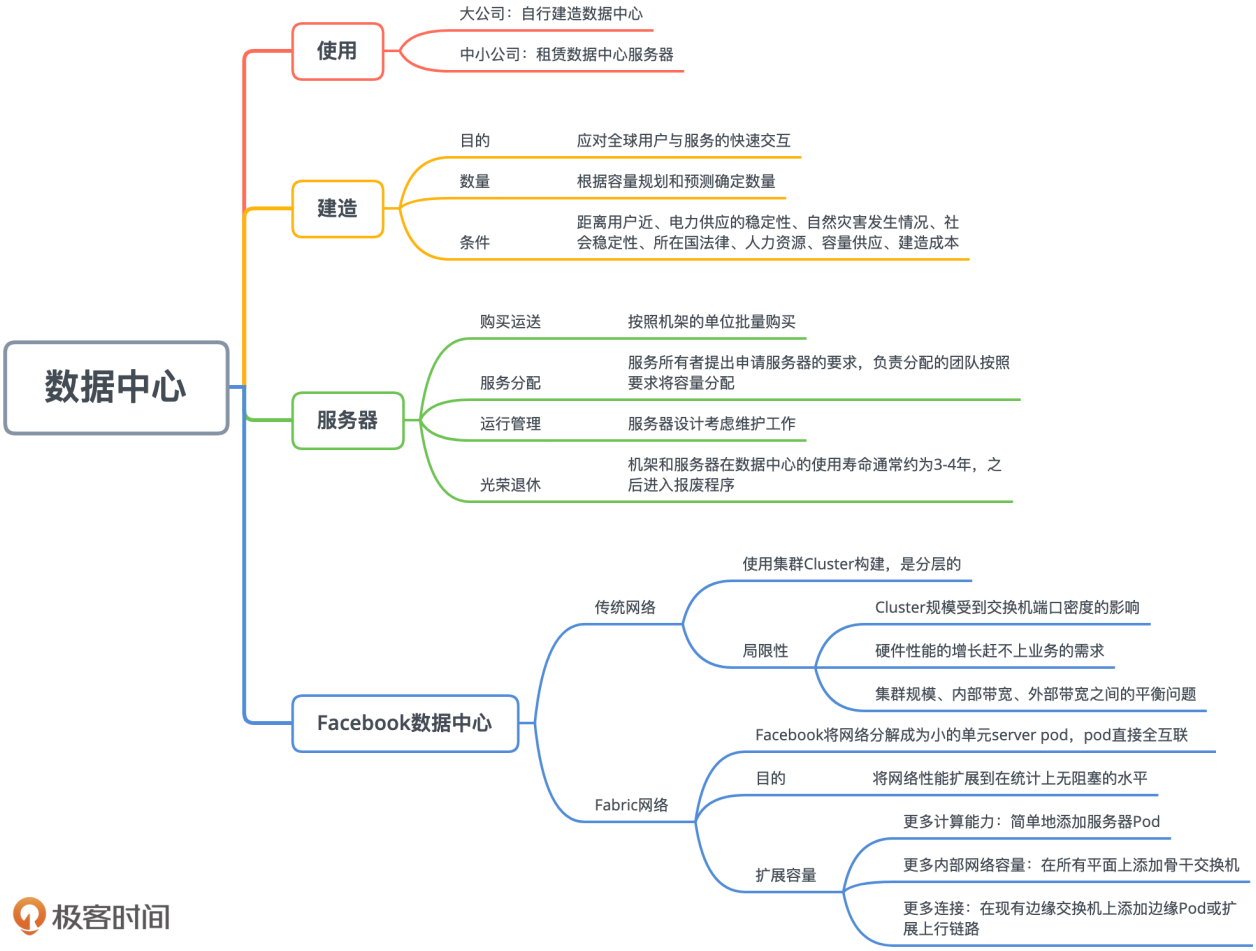
对于外部连接，我们用光纤网络配备了数量灵活的**边缘 Pod**。每个边缘 Pod，能够为骨干网和数据中心站点上的后端提供很多 Tbps 的带宽，并且可扩展到 100G 和更高的端口速度。

这种高度模块化的设计，使我们能够在一个简单统一的框架内，快速扩展任何维度的容量。

当我们需要更多计算能力时，就简单地添加服务器 Pod。当我们需要更多的内部网络容量时，就可以在所有平面上添加骨干交换机。当我们需要更多的连接时，我们可以在现有边缘交换机上添加边缘 Pod，或扩展上行链路。

总结

我们今天讨论了数据中心这个重要的容量载体，它的内部结构、网络设置、服务器的生命周期，以及规划数据中心的一些考虑因素。



数据中心可以说是服务器的“家”，也是我们的程序最终部署的地方。唐代诗人王建说：“今夜月明人尽望，不知秋思落谁家。”我们开发程序和部署服务，最好要了解数据中心的知识和架构，因为各种互联网服务，总归是需要数据中心的服务器和网络来支撑的。了解数据中心的配置，对我们服务的开发和部署是很有用的。

对一些大规模的服务，数据中心的网络或者服务器资源，可能会成为性能和业务发展的瓶颈，所以我们也需要不断优化，并提前规划数据中心。尤其是现在的互联网服务，往往有很大的数据量，所以数据中心网络的扩展性尤其重要。

思考题

你们公司的服务器一定也在数据中心里，这些数据中心的地理位置在哪里？地理因素对你的互联网服务的性能比如端到端延迟有什么影响？

你们公司用的数据中心是自己建造的还是租赁的？公司对数据中心的运营成本有什么要求和策略？

欢迎你在留言区分享自己的思考，与我和其他同学一起讨论，也欢迎你把文章分享给自己的朋友。

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇　30 | 服务器的管理和部署：工业界近几年有哪些发展趋势?

下一篇　32 | 服务的容量规划：怎样才能做到有备无患?

精选留言 (2)

写留言

钱晓明
2020-02-06

“四个独立的骨干交换机“平面”，每个平面，可以最多扩展 48 个独立设备”，这里的独立设备指的是骨干交换机吗？平面和边缘Pod是怎么连接的啊？

展开

作者回复: 数据中心的网络这块技术，我们公司内部也一直在不断地演化。你可以看看我们公司的一个博客，有稍微详细一点的内容：https://engineering.fb.com/production-engineering/introducing-data-center-fabric-the-next-generation-facebook-data-center-network/
更新更复杂的内容，你参考一下这个：https://engineering.fb.com/data-center-engineering/f16-minipack/ 希望能回答你的问题。

2

钱晓明

2020-02-06

老师，感觉网络是计算机或者云计算领域最复杂最难以理解的部分了，很重要的原因是对于大部分人来说没有机会接触到大规模的网络环境。今天说的数据中心网络内容很好，体现了一种设计思路，不知道这方面有没有更详细更具有实用价值的材料或者链接推荐呢？谢谢。

展开

作者回复: 谢谢评价！你说的很对。而且每个公司内部的数据中心都有自己独特的设计。像脸书，我们的网络也在不断演化，每个数据中心因为建造的时间不同，里面的架构也不一样。你可以看看我们比较新的一个相关博客： https://engineering.fb.com/data-center-engineering/f16-minipack/，有很多相关的知识。

1 1